CHAPTER I

PRELIMINARIES AND BASIC CONCEPTS

1.1 Introduction

The survival analysis has become an increasingly active and very important area of research. It is one of the core research methods used in many fields such as medicine, biology, epidemiology, demography and engineering. The survival analysis has several hundred years old at least as far as the publications of life tables by Halley (1693). The modern subject was studied by the presentation of product limit estimates by Kaplan-Meier (1958), generalization of Wilcoxon test by Gehan (1965) and the landmark paper by Cox (1972), which opened up the whole field of regression to survival analysis. It is used in various fields for analyzing data involving the duration between two events. It is also known as event history analysis, lifetime data analysis, reliability analysis or time to event analysis (Hosmer and Lemeshow, 1999). Applications of survival analysis now include time until onset of disease (Venkatesan, 1990, 2003), time until equipment failure, time until recurrence in a clinical study, time to death, time until infection and so on. In survival analysis the time variable referred is ‘survival time’ because it gives the time that an individual has survival over some follow-up period (Kalbfleisch and Prentice, 1980). Elandt-Johnson and Johnson (1980) describes applications in actuarial science and demography.

It is a collection of statistical methods used to address questions that have to do with whether and when an event of interest takes place. Precisely, it is “the analysis of data that correspond to the time from a well-defined time origin until the occurrence of some particular event or end-point” (Collett, 1994).
In medical research and treatment interventions, the researchers and practitioners closely monitor the timing of relapse of targeted problems because reducing the incidences and determining the timing of relapse are key measures of the interventions effectiveness. In all these instances, the timing of event occurrence is a key interest in research. Because the time-to-event data involve censoring, crucial statistical approaches cannot be readily employed. The censoring is a problem of incomplete data, that is, researchers cannot observe the entire histories of the targeted event for all study subjects and are unable to determine exactly the timing of event occurrence for a portion of study subjects. The fact that the event has not yet occurred for certain subjects at the time exclude the possibility that such an event will occur to them in the future. Because of censoring, all classical statistical approaches are not suitable for the analysis of survival data. Instead, researchers should use a special type of statistical tools, known as survival models and methods to analyze such data. Use a special type of regression model such as a discrete-time model, a Cox proportional hazards model or a parametric regression model to conduct multivariate analysis. These models are used to carry out the analysis based on standard Cox PH model (Venkatesan et al., 2011). The accelerated failure time models (AFT) are class of models used in the regression context to explore the survival data (Louis, 1981; Venkatesan and Raman, 2012; Ponnuraja and Venkatesan, 2010).

The studies in this field have previously focused on predicting the probability of response, survival or mean-lifetime, and comparing the survival distributions. In the recent years, the identification of risk and/or prognostic factors related to response, survival and the development of a disease has become equally important (Lee, 1992).

Some methods dealing with lifetime data were quite old like mortality tables, but many important methodological developments were largely achieved in the later half of the 20th century. Survival analysis concerns data on
time T to some event to happen; for example death, relapse and failure of a particular treatment.

1.2 Historical Developments

The earliest reference can be found in the 17th century (Graunt, 1662; Halley 1693; Dewitt, 1671). In the 18th century the Swiss mathematician Bernoulli used survival concepts for estimating the effects of smallpox inoculation on death rate and collected and evaluated mortality data of British troops during Crimean war.

Indeed, the first survival analysis was conducted approximately 350 years ago, when John Graunt derived the very first life table and published his famous paper “Natural and Political Observations Made Upon the Bills of Mortality” in 1662. The methods originate from biomedical interests in studying mortality, or patients’ survival times between the time of diagnosis of certain disease and death.

The life table reports have been published in the middle of 20th centuries (Berkson and Gage, 1950; Cutler and Ederer, 1958 and Gehan and Thomas, 1969). The product limit estimator by Kaplan-Meier (1958) is used for the estimation of the survival functions. The product limit estimator by Kaplan-Meier (1958) is used for the estimation of the survival functions. Its application pervades many areas from Biology and Medicine to Astronomy. The important non-parametric tests are Log-Rank test (Peto and Peto, 1972), Generalized Wilcoxon test (Breslow, 1972) and Tarone-Ware test (Tarone-Ware, 1977). The log rank test was developed by Peto and Peto (1972) based on the Savage (1956), Mantel’s (1966) and Altshuler (1970). Nelson (1969) introduced the hazard plotting used to select an appropriate parametric model for a data set. Different methods of comparison of empirical survival curves between groups were proposed by Mantel (1966), Gehan (1965) and Tarone and Ware (1977).
Peto et al., (1976) have published an outstanding review of some statistical methods related to clinical trials. Elandt-Johnson (1980) described applications in actuarial science and demography. Life tables are particularly suited for analyzing very large data sets (Young et al., 1999).

The Cox (1972, 1975) proportional hazard model for quantifying the effects covariates on the survival time and has had the most profound impact in the area. Aalen (1975) introduced an elegant martingale approach to survival analysis, where most of the methods can be cast within a unifying counting process framework. Aalen (1978) introduced the multiplicative intensity model as a generalization of the proportional hazards model. Using Martingale theory and other probability techniques, Slud and Wei (1982), Harrington and Flemming (1982) and Gu and Lai (1991) have developed the sequential analysis. Additive hazards model is explored as an alternative to the proportional hazards model by Cox and Oakes (1984), Thomas (1986) and Breslow and Day (1987). Multivariate failure time models have been studied by Dabrowska (1988), Prentice and Cai (1992) and van der Laan (1996) among others.

The failure data are well described by the exponential distribution by many authors (Davis, 1952; Crowder, Kimber, Smith and Sweeting, 1991; Cox and Oakes, 1984; Elandt-Johnson and Johnson, 1980; Gross and Clark, 1975; Lawless, 1982; Lee, 1992; Nelson, 1982, 1990). Collett (1994) presented parametric models at a level comparable to the topic at a slightly higher mathematical level. Major progress had been achieved and advance developments are in many other areas, together with the accelerated failure time model (Nelson and Hahn, 1972; Wei, 1992) multivariate failure time data (Mantel, 1983; Hougaard, 1986; Friedman, 1991), interval-censored data (So, 1994; Sun, 1997; Smith et al., 1997), dependent censoring (Wang, 2003), dynamic treatment regimens joint modeling of failure time and longitudinal data (Henderson,
The use of survival analysis in occupational epidemiology was illustrated by Liddle et al., (1977), Darby and Reissland (1981) and Breslow et al., (1983). Vaupel et al., (1979); and Clayton and Cuzick, (1985) studied methods to accommodate baseline heterogeneity in risk of developing tumors. The developments of censored rank tests were developed by Fleming et al., (1980) and Fleming and Harrington (1984). An initial attempt was made by Johnson et al., (1982), to assess the performance in a two covariate model for survival data. Petersen (1986) estimated full parametric survival models with time dependent covariates. Explore the effects on estimation if covariates are omitted from the proportional hazards model (Struthers and Kalbfleisch, 1986).

Andersen et al., (1993) discussed theoretical aspects of parametric regression models. A method for the regression analysis of interval-censored failure time data with focus on the comparison of failure time distributions among different treatment was proposed by Sun (1997).

As a special case gamma and log-normal frailty models have been studied by Therneau and Grambsch (2000). The inverse probability of censoring weighting technique (Cheng et al., 1995) which is used by Lin and Ying (1993) in the analysis of multivariate failure time data and also used to tackle the problems of dependent censoring by Lin et al., (1996). Frailty models are becoming increasing popular in multivariate survival analysis since they allow us to model the association between the individual survival times within subgroups. Frailty models were examined by Sahu et al., (1997), with a Weibull baseline hazard. Semi parametric approaches have also been examined by many authors Leemis et al., (1990).

The gamma distribution has been widely applied as a mixture distribution (Clayton, 1978; Vaupel et al., 1979; Oakes, 1982; Yashin and Iachine, 1995;
From a computational and analytical point of view the gamma distribution fits well to time-to-event data (Keiding et al., 1997), as it is easy to derive the closed form expressions of density and the hazard functions. The factor age influences more in heterogeneous among diseased cases in clinical trial (Vaupel et al., 1979). McGilchrist and Aisbett, (1991) described more than one event for obtaining individual and frailty common factors among recurrence times. With recent advances in Bayesian approaches, complex frailty models are now computationally feasible. Clayton (1991) and Sinha (1997) consider gamma process prior on the cumulative baseline hazard in the frailty model. Frailty models with piecewise exponential baseline hazards were discussed by many author (Sahu et al., 1997; Sinha and Dey 1997; Aslaindou et al., 1998). The frailty using Cox’s partial Likelihood was examined by Qiuo et al., (1999) and Sergent (1998). Bayesian approach to survival data analysis was studied by Chen et al., (1999). Survival analysis for recurrent event data and also determine the appropriate method for recurrent data using the key components Kelly and Lim (2000).

The problem of Bayesian variable selection for proportional hazards regression model for right censored data and proposed an MCMC based method to compute the posterior model probabilities (Ibrahim et al., 2000; Chen and Ibrahim, 2001). Yin and Ibrahim (2005) considered the multivariate failure time data and proposed a new class of Bayesian shared gamma frailty models by imposing the Box–Cox transformation on the hazard function. Bayesian semiparametric approach to inference was proposed by Michael et al., (2006). Bayesian analysis of a correlated frailty model and inverse Gaussian frailty was discussed by Kheiria et al., (2007).

The field of survival analysis experienced a tremendous growth during the later half of the 20th century after its applications in the field of cancer research. The
application of the methods of survival analysis has been greatly facilitated by the developments of computing power and statistical packages. Standard methods are now available virtually in all software and have popularized the methods. However, most of the newer methods, as evidenced by lack of use in the medical literature, are yet to be properly explored.

1.3 Basic Concepts

The random variable $T$ denotes the time to the event of our interest. $T$ is a positive random variable which has to be unambiguously defined that the start and end with the length of the time period in-between corresponding to $T$. All functions, unless stated otherwise, are defined over the interval $[0, t)$. The density function and distribution function are denoted by $f(t)$ and $F(t)$. The survival function is defined as

$$S(t) = P(T > t) = 1 - F(t) \quad \text{(1.1)}$$

where $F(t) = \int_0^\infty f(u)du$

The hazard function, $h(t)$ is the instantaneous rate of failure at time, and is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)} \quad \text{(1.2)}$$

The function $f(t)$, $F(t)$, $S(t)$, and $h(t)$ gives mathematically equivalent specifications of $T$. For censored data, specification of $f(t)$ or $F(t)$ is difficult where as $h(t)$ can be specified. Thus $S(t)$ is given by

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \quad \text{(1.3)}$$

1.4 Censoring

The special feature of survival analysis is censoring which is nothing but the incomplete observation of the failure times of individuals. This may be due to different
reasons like, individuals are still alive at the time of data analysis, they can no longer be traced or they may die due to reasons unrelated to the study. It distinguishes survival analysis from other branches of statistical methods. Survival data has certain feature as given below.

Failure time is positive valued and is generally not symmetrically distributed. It is usually very highly skewed and in the positive direction. The data cannot be assumed to follow a normal distribution. Many times, the usual Box-Cox transformations do not produce the desired results. A more satisfactory approach is to adopt an alternative distributional model for the original data. There are three major types of censoring as described below:

1.4.1 Type I Censoring

Let \( T_1, T_2, \ldots, T_n \) be \( n \) iid life times and \( t_c \) be fixed censoring time. Instead of observing \( T_1, T_2, \ldots, T_n \) the random variables of interest, we observe that

\[
Y_i = \begin{cases} 
T_i & \text{if } T_i \leq t_c \\
t_c & \text{if } T_i > t_c 
\end{cases}
\]

(1.4)

1.4.2 Type II Censoring

Let \( r < n \) be fixed and let \( T_1, T_2, \ldots, T_n \) be iid observations. Observation ceases after the \( r^{th} \) failure. Let \( T_{(1)} < T_{(2)} < \ldots < T_{(n)} \) be the ordered failure times, so the ordered observed sample is

\[
Y_{(1)} = T_{(1)}, Y_{(2)} = T_{(2)}, \ldots, Y_{(r)} = T_{(r)}, Y_{(r+1)} = T_{(r)}, \ldots, Y_{(n)} = T_{(r)}
\]

(1.5)

1.4.3 Random Censoring

Let \( C_1, C_2, \ldots, C_n \) be iid censoring times where \( C_i \) is associated with \( T_i \), \( i = 1 \) to \( n \). We can only observe \((Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)\) where

\[
Y_i = \min(T_i, C_i) = T_i \land C_i
\]

and
\[\delta_i = \begin{cases} 
1 & \text{if } T_i \text{ is uncensored} \\
0 & \text{if } T_i \text{ is censored} 
\end{cases} \] (1.6)

Notice that \(Y_1, Y_2, \ldots, Y_n\) are iid with some degrees of freedom and \(\delta_i\) is the indicator variable. While observing the response or survival times, censoring occurs in the following form (a) Lost to follow-up (b) Dropout (c) Termination of the study.

A key assumption is made for all the models, is that survival times and the censored times are independent. Most of the methods of survival analysis may not be accurate if this assumption is violated.

1.4.4 Right and Left Censoring

Right censoring, which is more common form, occurs when the exact survival time is not known. All that is known is that the exact survival time exceeds the recorded value. This occurs when there is a defined time \((t=0)\) where the observation of time is started for all subjects involved in the study. A right-censored subject's time terminates before the outcome of interest is observed.

Left censoring is not common in clinical trial; an observation is left censored if the event of interest has already occurred when observation of time begins. In other terminology left censored data can occur when a subject’s survival time is incomplete at the beginning of the follow up period. A best example for left censored data in a clinical trial is, among the subjects with HIV infection, knowing the subject is HIV positive only when the patient having been undergone some test procedure which is for confirming the disease present or not, but we may not know when the viral infection entered in his body.

1.5 Maximum Likelihood Estimation
We assume the random censoring model (Miller, 1981).

The pair \((Y_i, \delta_i)\) has likelihood.

\[
L(Y_i, \delta_i) = \begin{cases} 
  f(y_i) & \text{if } \delta_i = 1 \text{ (uncensored)} \\
  S(y_i) & \text{if } \delta_i = 0 \text{ (censored)} 
\end{cases}
\]

\[
= f(y_i)^{\delta_i} S(y_i)^{1-\delta_i},
\]

and the likelihood of the full sample is

\[
L = L(Y_1, ..., y_n; \delta_1, ..., \delta_n) = \prod_{i=1}^{n} L(Y_i, \delta_i)
\]

\[
= \left( \prod_{u} f(y_i) \right) \left( \prod_{c} S(y_i) \right)
\]

(1.8)

where \( \prod_u \left( \prod_c \right) \) denotes a product over the uncensored (censored) observation.

The complete likelihood under random censoring is

\[
L(y_i, \delta_i) = \begin{cases} 
  f(y_i) (1 - G(y_i)) & \text{if } \delta_i = 1 \\
  g(y_i) S(y_i) & \text{if } \delta_i = 0 
\end{cases}
\]

(1.9)

where \( G(y_i) \) is the cumulative distribution function for censoring.

\[
L = \left( \prod_{u} f(y_i) \right) \left( \prod_{c} S(y_i) \right) \left( \prod_{c} g(y_i) \right) \left( \prod_{u} (1 - G(y_i)) \right)
\]

(1.10)

But under the assumption that the censoring time has no connection to the survival time, the last two products \( \prod_c g(y_i) \) and \( \prod_u (1 - G(y_i)) \) do not involve the unknown life time parameters, so these two products can be treated like constants when maximizing \( L \).

Let \( \theta = (\theta_1, ..., \theta_p)' \) be the vector of parameters. Finding \( \text{max } L(\theta) \) is equivalent to finding the solution \( \hat{\theta} \) to the likelihood equations.

\[
0 = \frac{\partial}{\partial \theta_j} \log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log L(y_i, \delta_i)
\]
\[ 0 = \sum_{u} \frac{\partial}{\partial \theta_j} \log f_u(y_i) + \sum_{v} \frac{\partial}{\partial \theta_j} \log S_u(y_i), \quad j = 1, \ldots, p \]  \hspace{1cm} (1.11)

Using Newton-Raphson and Method of Scoring we could solve the above equation Gross and Clark (1975), Kalbfleisch and Prentice (1980).

Denote

\[ L_i(\theta) = L_0(y_i, \delta_i), \quad i = 1, \ldots, n \]

and

\[ \frac{\partial}{\partial \theta} \log L(\theta) = \frac{\partial}{\partial \theta_1} \log L(\theta), \ldots, \frac{\partial}{\partial \theta_p} \log L(\theta), \]

\[ \frac{\partial^2}{\partial \theta^2} \log L(\theta) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1 \partial \theta_1} \log L(\theta), & \ldots, & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \log L(\theta) \\ \ldots & \ldots & \ldots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1} \log L(\theta), & \ldots, & \frac{\partial^2}{\partial \theta_p \partial \theta_p} \log L(\theta) \end{pmatrix} \]  \hspace{1cm} (1.12)

Then the likelihood equations are

\[ 0 = \sum_{i} \frac{\partial}{\partial \theta_j} \log L_i(\theta), \quad j = 1, \ldots, p; \quad \text{ (or) } \]

\[ \hat{\theta} = \frac{\partial}{\partial \theta} \log L(\theta) \]  \hspace{1cm} (1.13)

Assume \( \hat{\theta}^0 = (\hat{\theta}_1^0, \ldots, \hat{\theta}_p^0) \) is an initial guess at the solution. Expand about \( \hat{\theta}^0 \):

\[ 0 = \sum_{i} \frac{\partial}{\partial \theta_j} \log L_i(\hat{\theta}) + \sum_{j} \frac{\partial}{\partial \theta_j} \log L_i(\hat{\theta}^0) + \sum_{k} \frac{\partial^2}{\partial \theta_k \partial \theta_j} \log L_i(\hat{\theta}^0) + \ldots \quad j = 1, \ldots, p \quad \text{ (or) } \]
\[
0 = \frac{\partial}{\partial \hat{\theta}} \log L(\hat{\theta}) = \frac{\partial}{\partial \hat{\theta}} \log L(\hat{\theta}^0) + \frac{\partial^2}{\partial \hat{\theta}^2} \log L(\hat{\theta}^0)(\hat{\theta} - \hat{\theta}^0) + ... \quad (1.14)
\]

Let \( \hat{\theta}^0 \) be the solution ignoring second order and higher terms:

\[
\hat{\theta}^0 = \theta^0 + \left(-\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta}^0)\right)^{-1} \frac{\partial}{\partial \theta} \log L(\hat{\theta}^0) \quad (1.15)
\]

The vector \( \frac{\partial}{\partial \theta} \log L(\hat{\theta}^0) \) is called the score vector at \( \hat{\theta}^0 \), and the matrix

\[
i(\hat{\theta}^0) = -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta}^0) \quad (1.16)
\]

is called the sample information matrix at \( \hat{\theta}^0 \). Notice that

\[
E[i(\hat{\theta})] = \left[-E \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\hat{\theta})\right] = I(\hat{\theta}) \quad (1.17)
\]

which is the Fisher information. We point out that \( I(\hat{\theta}) \) is the Fisher information of the entire sample:

\[
I(\hat{\theta}) = \sum_{i=1}^{n} I_i(\hat{\theta}) = n I_i(\hat{\theta}) \quad (1.18)
\]

where \( I_i(\hat{\theta}) \) is the Fisher information of the \( i^{th} \) observation

The iteration scheme using (1.19) is called the Newton-Raphson method. Replacing the sample information in (1.19) by the Fisher information gives

\[
\hat{\theta}^i = \hat{\theta}^0 + I^{-1}(\hat{\theta}^0) \frac{\partial}{\partial \theta} \log L(\hat{\theta}^0) \quad (1.19)
\]
and the iteration scheme using the above equation is called the Method of Scoring. While that equation might produce improved convergence in some instances, it may not be possible, particularly if censoring is present, to figure out $\hat{I}(\hat{\theta})$ for use in (1.20).

1.5.1 Confidence Intervals

Confidence intervals and tests for random and Type I censoring under smoothness conditions,

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta))$$  \hspace{1cm} (1.20)

Usually for Type II censoring, this result also holds, but the proofs are different. (The notation $\hat{a}$ denotes, is asymptotically distributed as)

For testing $H_0 : \theta = \theta^0$ or constructing confidence intervals, we have three procedures.

(i) Wald

$$\left(\hat{\theta} - \theta^0\right)\sim L(\theta^0)\left(\hat{\theta} - \theta^0\right) \sim \chi_p^2 \text{ under } H_0.$$ We can alternatively substitute

$$\hat{I} \left( \hat{\theta} \right) \text{ for } I(\theta^0).$$

(ii) Neyman-Pearson / Wilks likelihood ratio

$$-2 \log \frac{L(\theta^0)}{L(\hat{\theta})} \sim \chi_p^2 \text{ under } H_0.$$ (iii) Rao

$$\frac{\partial}{\partial \theta} \log L(\theta^0) I^{-1}(\theta^0) \frac{\partial}{\partial \theta} \log L(\theta^0) \sim \chi_p^2 \text{ under } H_0.$$
Notice that Rao’s method does not use the Maximum Likelihood Estimator (MLE), so no iterative calculation is necessary. However, in addition to tests, we usually want estimates and confidence intervals, so we should need to calculate \( \hat{\theta} \) anyway. Once we have \( \hat{\theta} \) and \( I(\theta^0) \), the Wald method is easy.

Under censoring we may need to replace \( I(\theta) \) with \( i(\theta) \) because calculation of \( I(\theta) \) is usually difficult. Also Efron and Hinkely (1978) suggest that using \( i(\theta) \) is better than using \( I(\theta) \) for confidence intervals even if \( I(\theta) \) can be calculated. There is not universal agreement on this, however.

### 1.6 Non-Parametric Maximum Likelihood Estimation

In the analysis of survival data if an uncensored sample of \( n \) distinct failure times is observed from a homogenous population, the sample survival function is a step function decreasing by 1 immediately following each observed failure time. A simple adjustment is made if any ties present in the data. However survival data very often involves right censoring and a methodology for handling this is required.

Let \( t_1 < t_2 < \ldots < t_k \) represent the observed failure times in a sample of size \( n_0 \) from a homogenous population with survival function \( S(t) \), suppose that \( d_j \) items fail at \( t_j \) (\( j = 1, \ldots, k \)) and \( m_j \) items are censored in the interval \( (t_j, t_{j+1}) \) at times \( t = 1, \ldots, m_{jt} \) (\( j = 0, \ldots, k \)) where \( t_0 = 0 \) and \( t_{k+1} = \infty \). The probability of failure at \( t_j \) is

\[
S(t_j) - S(t_j + 0)
\]

where \( S(t_j + 0) = \lim_{x \to 0} S(t_j + x) \), \( j = 1, \ldots, k \) we assume that the contribution to the likelihood of a survival time censored at \( t_{j1} \) is
\[ P(T > t_{j1}) = S(t_{j1} + 0) \quad (1.21) \]

In effect we are assuming that the observed censoring time \( t \) tells that the unobserved failure time is greater than \( t_{j1} \) and thus we obtain.

\[
L = \prod_{j=1}^{m_j} \left( S(t_j) - S(t_j + 0) \right)^{d_j} \prod_{j=1}^{m_j} S(t_j + 0) \quad (1.22)
\]

which given data, can be viewed as a likelihood function on the space of survival function \( S(t) \) that maximizes \( L \).

This definition of MLE is the generalization of the usual concept used in parametric models.

Clearly \( S(t) \) is discontinuous at the observed failure times since otherwise \( L = 0 \) further, subject to \( t_{j1} > t_j \), \( S(t_{j1} + 0) \) is maximized by taking.

\[
S(t_{j1} + 0) = S(t_{j1} + 0), \quad j = 1, \ldots, k \quad \text{and} \quad l=1, \ldots, m_j \quad (1.23)
\]

and

\[
S(t_{il}) = 1, i = 1, \ldots, m. \quad (1.24)
\]

The function \( S(t) \) is then a discrete survival function with hazard components \( h_1, \ldots, h_k \) at \( t_1, \ldots, t_k \) respectively. Thus

\[
S(t_j) = \prod_{i=1}^{j-1} \left( 1 - \hat{h}_i \right) \quad (1.25)
\]

\[
S(t_j + 0) = \prod_{i=1}^{j} \left( 1 - \hat{h}_i \right) \]

where the \( \hat{h}_i \) are chosen to maximize the function

\[
\prod_{j=1}^{k} \left[ \prod_{i=1}^{j-1} (1 - h_i)^{d_j} \right] h_j^{d_j} \left[ \prod_{i=1}^{j} (1 - h_i)^{m_i} \right] = \prod_{j=1}^{k} h_j^{d_j} (1 - h_j)^{n_j - d_j} \quad (1.26)
\]

Obtained by substitution of the above in \( L \). Clearly \( \hat{h}_j = d_j / n_j \), \( j = 1, \ldots, k \) and the product limit of the survival function is
\[ S(t) = \prod_{j:t_j \leq t} \frac{(n_j - d_j)}{n_j} \quad (1.27) \]

In obtaining the product limit estimate, we are in effect making the conditional probability of failure at each \( t_j \) agree exactly with the observed conditional relative frequency of failure at \( t_j \) given by \( d_j / n_j \).

A slightly problematic point is that \( S(t) \) never reduces to zero if \( m > 0 \), for this reason \( S(t) \) is usually taken to be undefined for \( t > t_{km} \).

The non-parametric maximum likelihood estimate needs a word of explanation since measure theoretic difficulties arise in interpreting \( L \). If one considers the data to have arisen from a continuous distribution with a small measurement error associated with each observed failure time, the first factor of \( L \) becomes \( \prod [(t_j + \delta_j) - (t_j - \delta_j)] \) where the interval \([t_j - \delta_j), (t_j + \delta_j)\] = \( A_j \) is defined by the precision of measurement.

In order to avoid difficulties with censoring times tied with failure times, we adopt the convention of moving such censoring times a small amount to the right. Consider now \( \delta_j \rightarrow 0 \) and suppose that \( \delta_j \) is small enough that all censoring points are excluded from \( A_j \) with \( \delta_j \) so chosen, \( S \) is a maximum likelihood estimate and converges to the product limit estimate as \( \delta_j \rightarrow 0 \).

The estimate \( \hat{S}(t) \) is the direct generalization of the sample survival function for censored data. It was first derived by Kaplan and Meier (1958) and as a consequence, is often referred to as the Kaplan-Meier estimate. The asymptotic variance of \( \hat{S}(t) \) can be obtained using Greenwoods formula as

\[ \text{var} \hat{S}(t) = S(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{\{n_j(n_j - d_j)\}} \quad (1.28) \]
1.7 Recurrent Events Data

In clinical trials events may occur more than once over the follow-up time for a subject. It’s known as ‘recurrent events’ (Nelson, 2003). It is usually possible to study just the time until the first event, it may be useful to incorporate subsequent events to increase the information available. Modeling this kind of data can be carried out using Cox-PH model (4 types) with the re-constructed data layout, so that each subject has a line of data corresponding to each recurrent event (Lawless and Nadeau, 1995). A variation of this approach uses a stratified Cox-PH model (Kleinbaum, 1996), which stratifies on the order in which the recurrent events occur.

It considers adjusting the variances of estimated model coefficients for the likely correlation among recurrent events within the same subject. This adjusted variance estimates are known has ‘robust variance estimates’ (Clayton, 1994).

The four types of modeling the recurrent events analysis are: counting process, conditional A, conditional B and marginal. These models are used to carry out the analysis based on standard Cox PH model (Venkatesan et al., 2011). A stratified Cox model or an extended Cox model would need to be used if one or more time-independent variables did not satisfy the PH assumption. Also, an extended Cox model would be required if inherently time-dependent variables were considered.

1.8 Competing Risks Analysis

Problems involving competing risks (Crowder, 2001) are common in medical applications. In such problems, there are K competing causes of failure that may occur. When one of the competing causes occurs the occurrence of the other causes are not possible. One observes for each unit simply a failure time and a cause of failure.

In medical applications, competing risks are found in many situations (Fine and Gray, 1999). A common example is the analysis of cause of death data. In clinical
studies common competing risks are relapse and death in remission (Lunn and McNeil, 1995). Interest is often on estimating the rate of occurrence of the competing risks, comparing these rates between treatment groups and modeling the effect of covariates on the rate of occurrence of the competing risks.

If only one failure type is of primary interest, then the analysis might be restricted to estimating hazards or hazard ratios for that type only. To describe this method mathematically, the cause-specific hazard function given below. The random variable $T_c$ denotes the time-to-failure from event type $c$. Thus, $h_c(t)$ gives the instantaneous failure rate at time $t$ for event type $c$, given not failing from event $c$ by time $t$.

Using a Cox PH model that considers predictors $X=(X_1,X_2,…,X_p)$, the cause-specific hazard model for event type $c$ has the form shown at the left. Note that $\beta_{ic}$, the regression coefficient for the $i^{th}$ predictor, is subscripted by $c$ to indicate that the effects of the predictors may be different for different event types.

1.8.1 Cause-Specific Hazard Function

$$h_c(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T_c < t + \Delta t | T_c \geq t)}{\Delta t}$$

where $T_c$ - time-to-failure from event $c$; $c = 1, 2,…, C$

1.8.2 Cox PH Cause-Specific Model

$$h_c(t, X) = h_{0c}(t) \exp \left[ \sum_{i=1}^{p} \beta_{ic} X_i \right]$$

where $c = 1,…,C$ ; $\beta_{ic}$ allows effect of $X_i$ to differ by event-type
Modeling competing risks survival data is usually carried out using a Cox model, parametric survival models or models that use the cumulative incidence.

### 1.9 Frailty Models

A frailty model is a multiplicative hazard model consisting of three components: a frailty (random effect), a baseline hazard function (parametric or nonparametric), and a term modeling the influence of observed covariates (fixed effects). Gamma distribution is the most often applied frailty distribution because frailties appearing in conditional likelihood can be integrated out, giving simple expressions of unconditional likelihood. Then, maximization of unconditional likelihood can be used for estimation. Here, the interpretation of frailty is as unobserved heterogeneity due to non observed covariates. The shared frailty model has been discussed in detail by other authors (Hougaard, 2000; Therneau and Grambsch, 2000). Shared frailty models are an important tool for analyzing multivariate (clustered) survival data. In general a frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals. That is, there may be factors other than the measured covariates that significantly affect the distribution of survival time. When these factors are present in the usual normal errors linear model setting, they are called random effects. Survival models incorporating such factors are called Frailty models.

The notion of frailty provides a convenient way to introduce random effect, association and unobserved heterogeneity into models for survival data. The term frailty itself was introduced by Vaupel et al., (1979) in univariate survival models and the model was applied by Clayton (1978) to multivariate survival data in a seminal paper on chronic diseases incidence in families. In essence, the frailty concept goes back to work (Greenwood and Yule, 1920) on accident proneness. Frailty models are
extensions of proportional hazard model which is best known as the Cox’s model Cox (1972), the most popular model in survival analysis. Normally, in most clinical applications, survival analysis implicitly assumes a homogeneous population to be studied. The frailty approach is a statistical modeling concept which aims to account for heterogeneity, caused by unmeasured factors. In statistical terms, a frailty model is a random effect model for time-to-event data, where the random effect (the frailty) has a multiplicative effect on the baseline hazard function. There are two broad classes of frailty models are models with a univariate survival time as end point and models which describe multivariate survival end points (e.g., competing risks, recurrence of events in the same individual, occurrence of disease in relatives).

One important problem in the area of frailty models is the choice of the frailty distribution. The frailty distributions most often applied are the gamma distribution (Clayton, 1978; Vaupel et al., 1979), the Positive Stable distribution Hougaard (1986b), a three-parameter distribution namely Power Variance Function family of frailty model (PVF) (Hougaard, 1986a), the Compound Poisson distribution (Aalen, 1988, 1992) and the log-normal distribution (McGilchrist and Aisbett, 1991).

Univariate frailty models are widely applied. A few examples are listed here. Aalen and Tretli (1999) applied the Compound Poisson distribution to testicular cancer data. The idea of the model is that a subgroup of men is particularly susceptible to cancer, which results in selection over time.

1.10 Bayesian Survival Analysis

Bayesian analysis of survival data has received much recent advances in computational and modeling techniques. With the use of Bayesian inference particularly Gibbs sampler and other Markov Chain Monte Carlo techniques, fitting
complex models is fairly straightforward, and the availability of software Bayesian Using Gibbs Sampler (BUGS) (Spiegelhalter et al., 2003) eases the implementation to a great extent. MCMC sampling enables us to make exact inference for any sample size without resorting to asymptotic calculations.

With the specification of $h(t), S(t), f(t)$ and $F(t)$ The hazard function $h(t)$ has the properties

$$h(t) \geq 0 \text{ and } \int_0^\infty h(t) dt = \infty \quad (1.31)$$

and

$$f(t) = h(t) \exp \left( - \int_0^t h(u) du \right) \quad (1.32)$$

The hazard function depends in general on both time and covariates, of which some may be time dependent, the proportional hazard model Cox (1972) separates these components by specifying that the hazard at time $t$ for an individual whose covariate vector is $x$ is given by

$$h(t \mid x) = h_0(t) \exp \{ G(x, \beta) \} \quad (1.33)$$

where $h_0(t)$ is called the baseline hazard function and $\beta$ is a vector of regression coefficients and the exponential part must be positive. Moreover this model implies that the ratio of hazards for two individuals is constant over time. It is convenient to assume that the effect on the covariates is multiplicative, the hazard function

$$h(t \mid x) = h_0(t) \exp(x', \beta) \quad (1.34)$$

This model implies that the ratio of hazards for two individuals depends on the difference between their linear predictors at any time. Where, $x' \beta = \eta$, is the linear predictor.

1.11 Aim and Objectives
In this thesis we have concentrated on the semi-parametric survival regression models, their extension and competing risk models with applications to clinical trials.

The broad objectives of the research work are:

- To collect and unify the recent advances in survival and competing risk models.
- To study new approaches for treatment-patient interactions using frailty models.
- To propose and investigate new models for the recurrent survival events data analysis.
- To explore the use of competing risk models for estimating the cause specific hazard rates for toxicity and relapse in HIV/TB trials.

1.12 Database and Software

The data have been collected from National Institute for research in Tuberculosis, Chennai. The two different data was collected. The first data consists of 336 HIV infected Tuberculosis patients admitted in clinical trial who were treated with two treatments of six months duration and nine months duration. The event of interest is death and follow-up period is 5 years.

The second database consists of 155 multi drug resistant tuberculosis patients admitted in a clinical trial at National Institute for Research in Tuberculosis, Chennai. The event of interest is sputum culture conversion from positive to negative in every occasion. The analysis was carried out using the software’s, WinBUGS, R, Stata, SAS and SPSS.